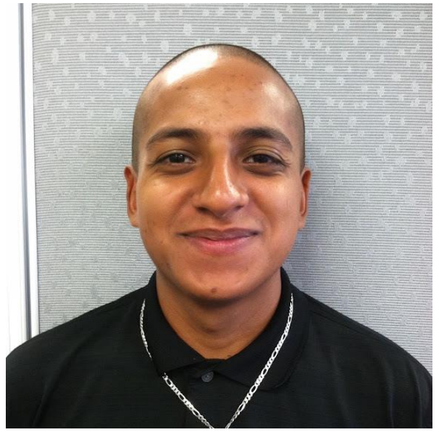
St. Jude Children's Research Hospital®
Finding cures. Saving children.

Introduction to sc/snRNA-seq & visualization with R shiny applications

Sharon Freshour, PhD
Bioinformatics Research Scientist
Bioinformatics Core
Department of Developmental Neurobiology
St. Jude Children's Research Hospital
September 18th, 2024

The DNB Bioinformatics Core Team



Cody Ramirez, Ph.D.

Senior Bioinformatics Research Scientist
Core Director
Boston, Massachusetts

Antonia Chroni, Ph.D.

Senior Bioinformatics Research Scientist
New York, New York

Asha Jannu, Ph.D.

Bioinformatics Research Scientist
Indianapolis, Indiana

Sharon Freshour, Ph.D.

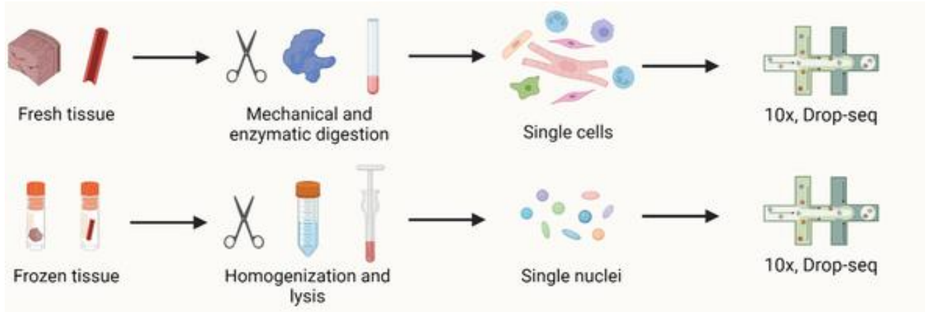
Bioinformatics Research Scientist
St. Louis, Missouri

Intro to sc/snRNA-seq Workshop Overview

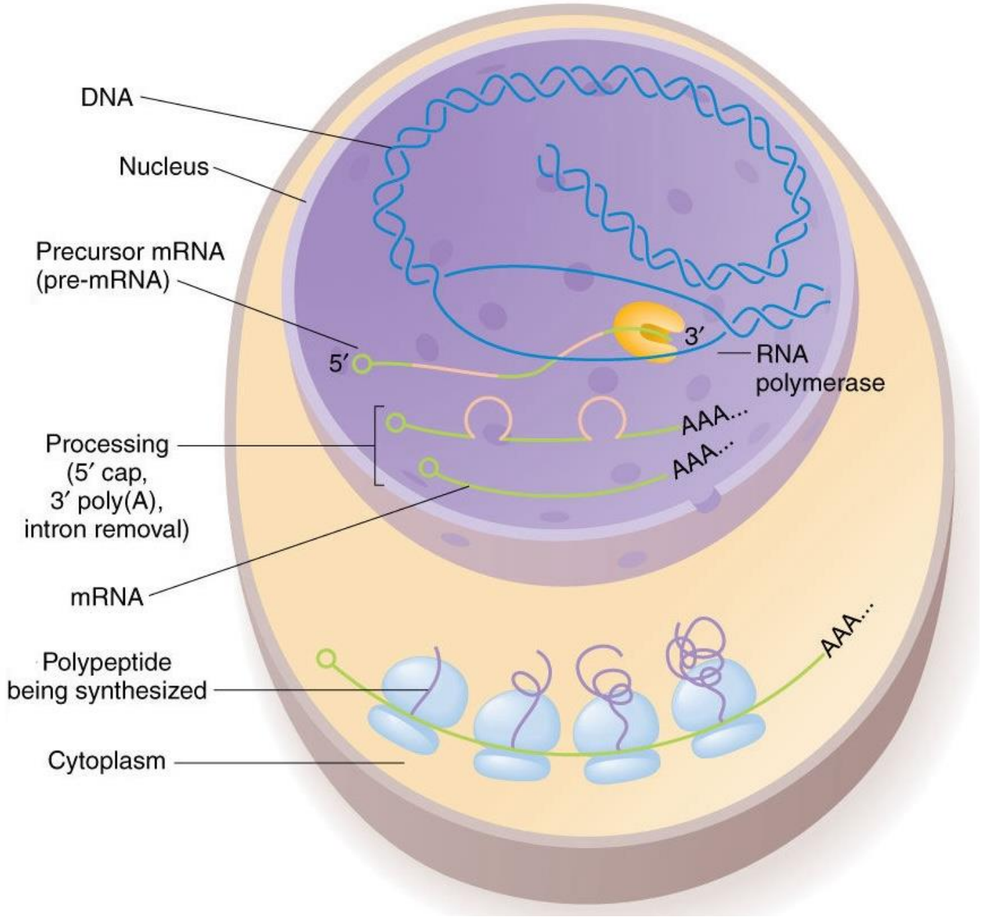
- Single cell (scRNA-seq) vs single nuclear (snRNA-seq) RNA sequencing
- sc/snRNA-seq library preparation
- Raw sequencing QC (FastQC)
- Alignment QC (CellRanger)
- Cell-level filtering (Seurat)
- Short break for everyone to get their lunch
- Visualization with R Shiny

scRNA-seq versus snRNA-seq

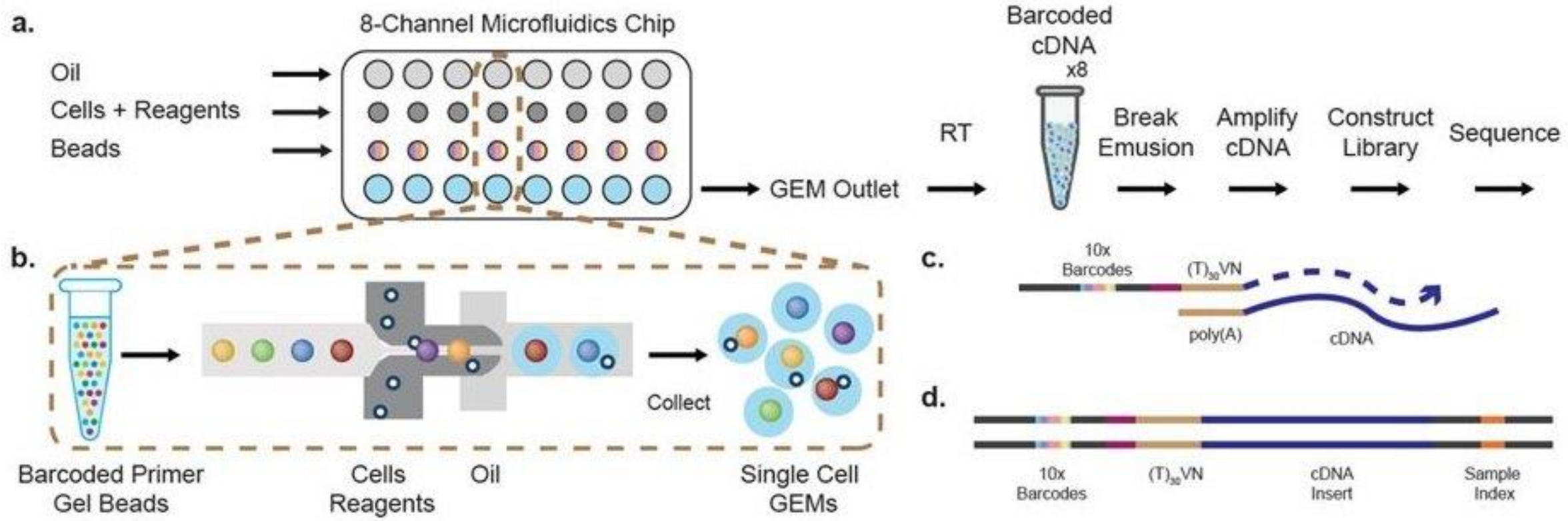
	scRNA-seq	snRNA-seq
Portion of cell:	Entire cell	Nuclei
Transcripts measured:	Cytoplasmic + nuclear	Nuclear
Pre-mRNA %:	Lower	Higher
mRNA %:	Higher	Lower
Exonic %:	Higher	Lower
Intronic %:	Lower	Higher
More appropriate for:	Fresh	Preserved or isolation difficulties



Xu, Xinjie, et al. DOI: 10.1007/s00395-022-00972-1.



10x Genomics sc/snRNA-seq platform

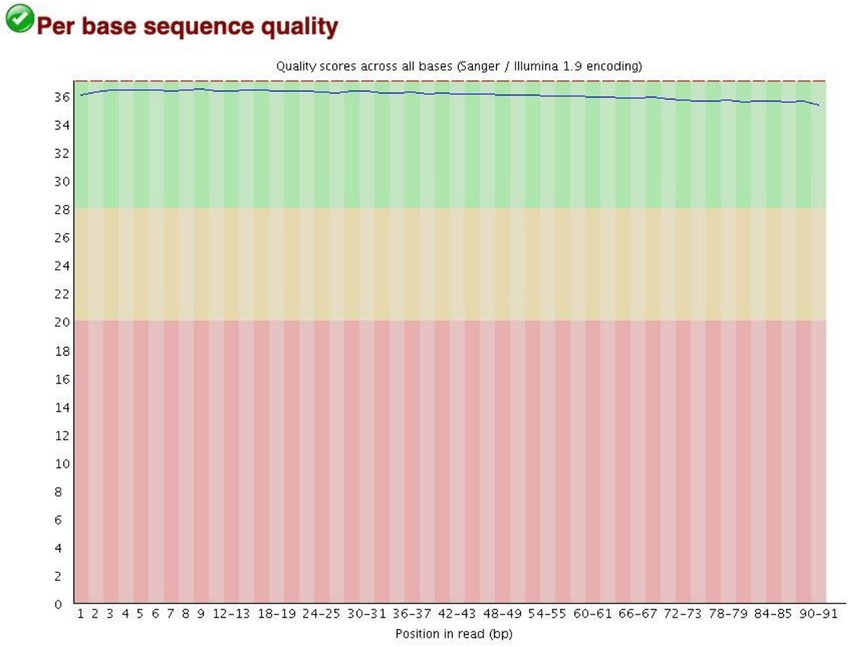


Zheng, G., Terry, J., Belgrader, P. et al. DOI: 10.1038/ncomms14049

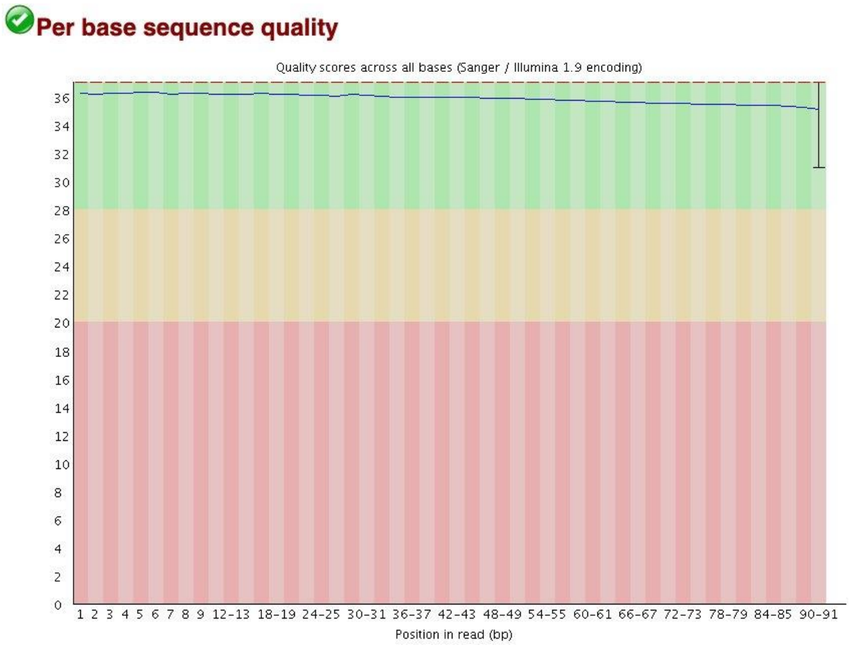
Quality control of raw sequencing data with FastQC

Per base sequence quality indicates the sequencing quality scores at each base position

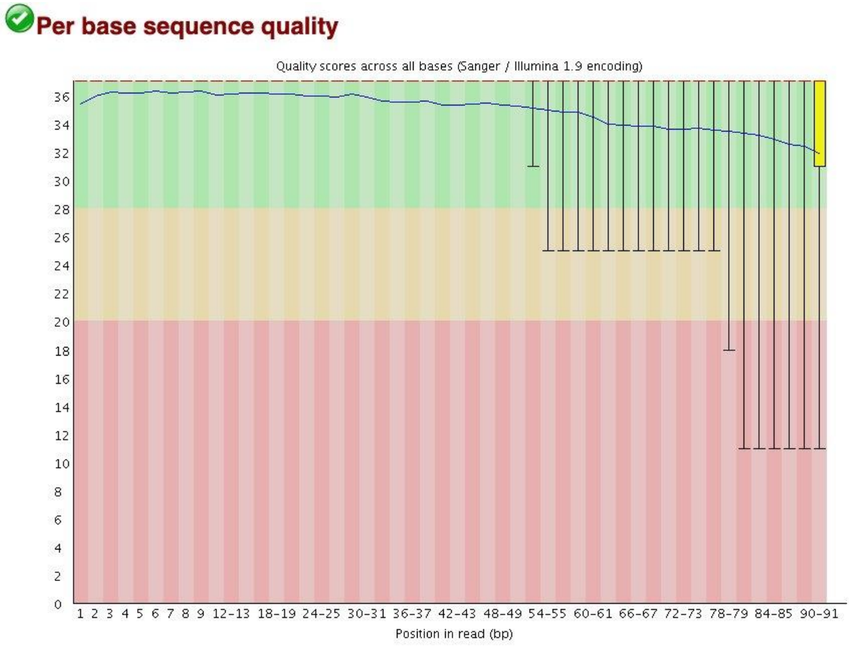
Gold standard
10x Genomics scRNA-seq



Salvageable?
Patient sample snRNA-seq

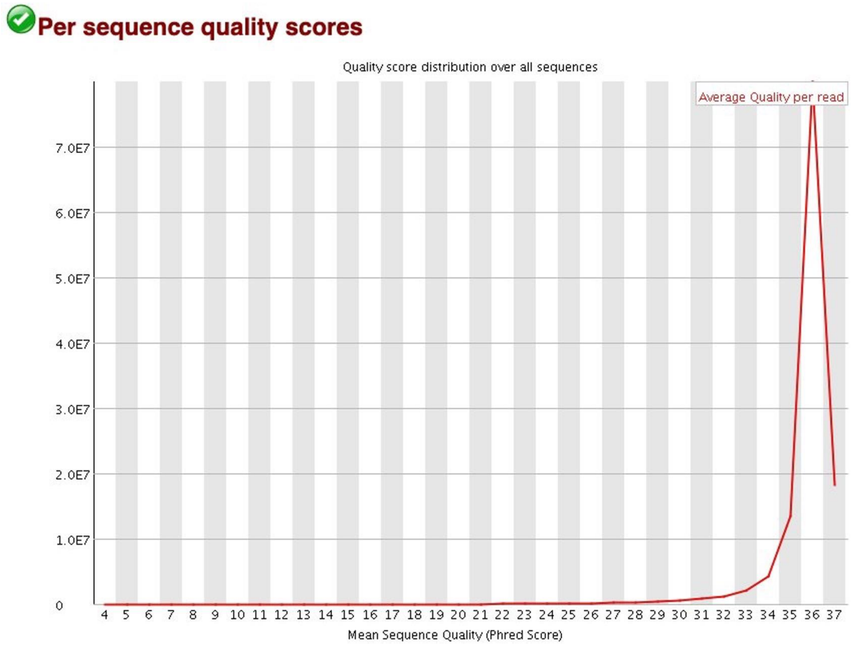


Fail
Organoid sample snRNA-seq

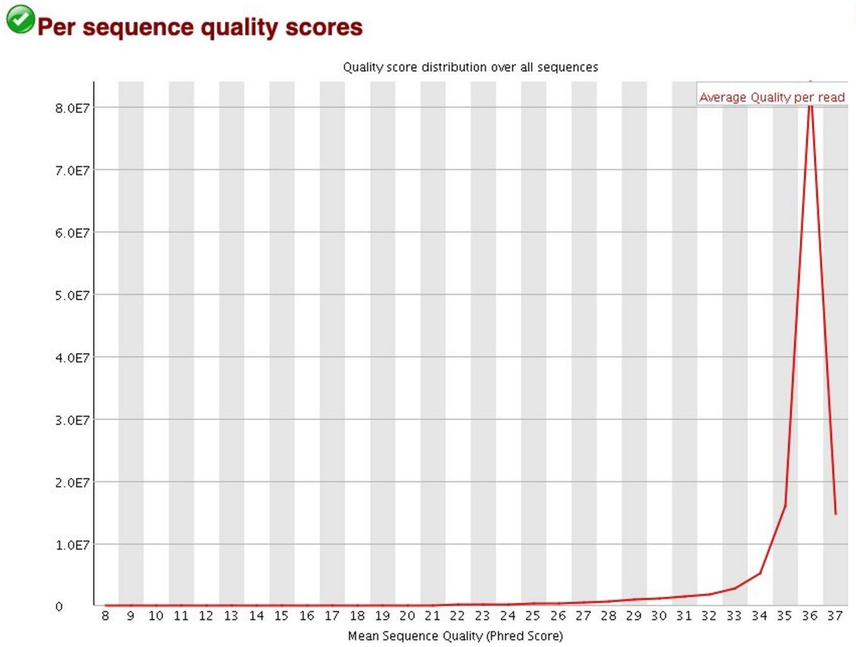


Per sequence quality scores indicate if sequences have poor quality across sequencing run

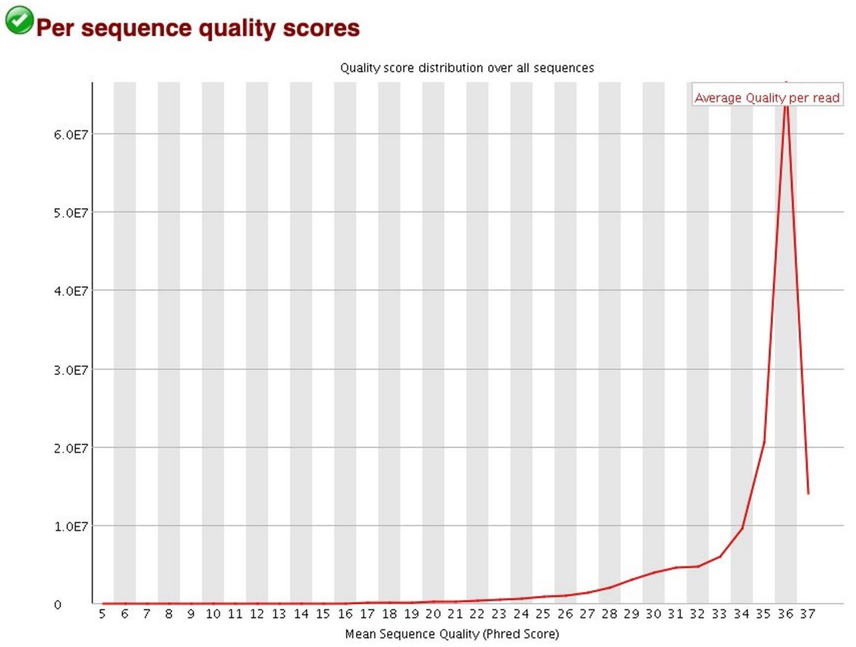
Gold standard
10x Genomics scRNA-seq



Salvageable?
Patient sample snRNA-seq

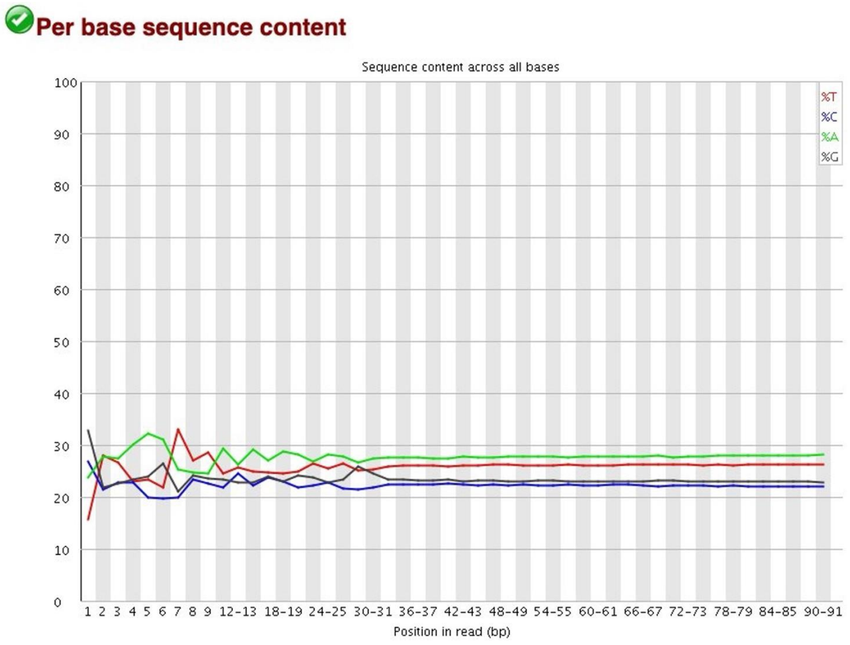


Fail
Organoid sample snRNA-seq

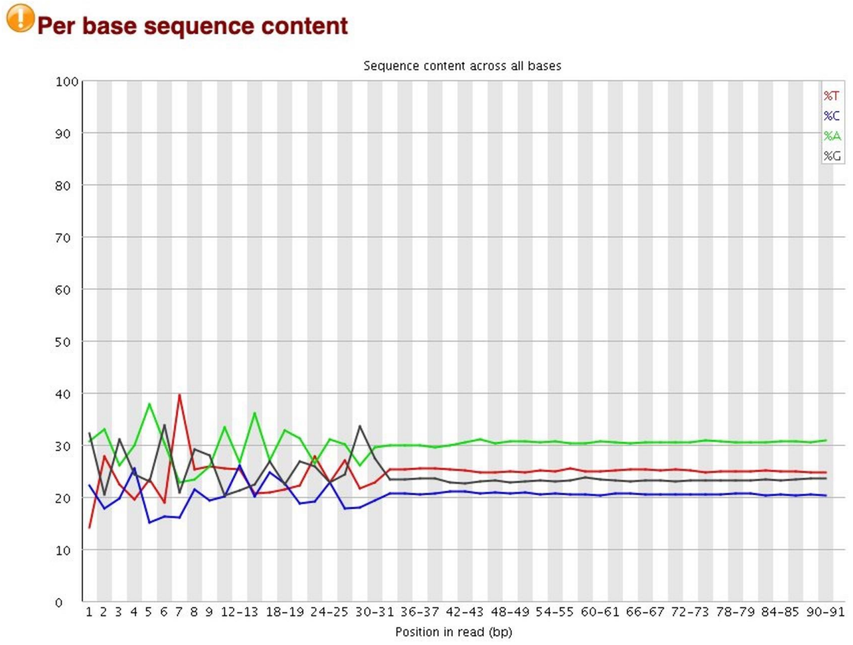


Per base sequence content indicates if there is (unexpected) bias in the proportion of bases called

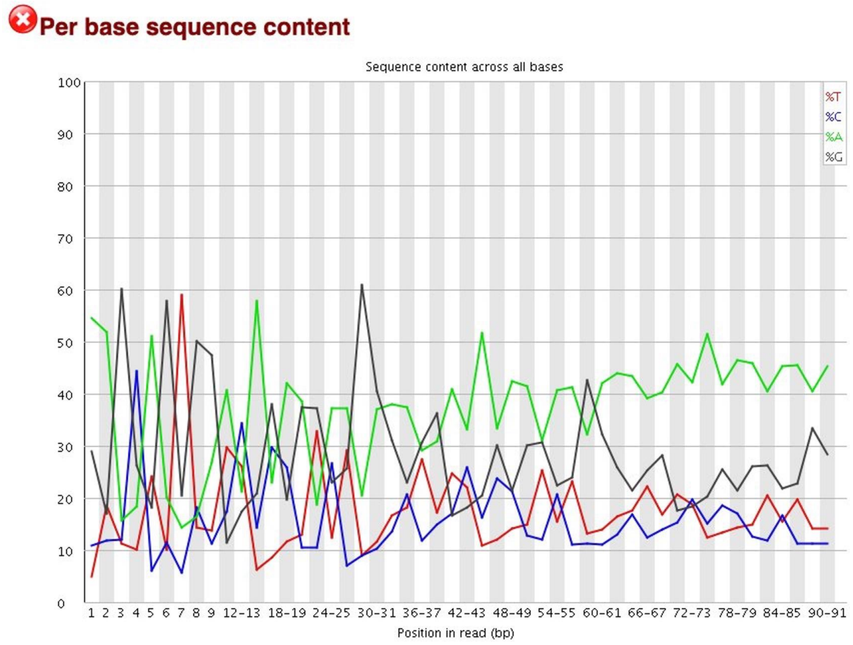
Gold standard
10x Genomics scRNA-seq



Salvageable?
Patient sample snRNA-seq

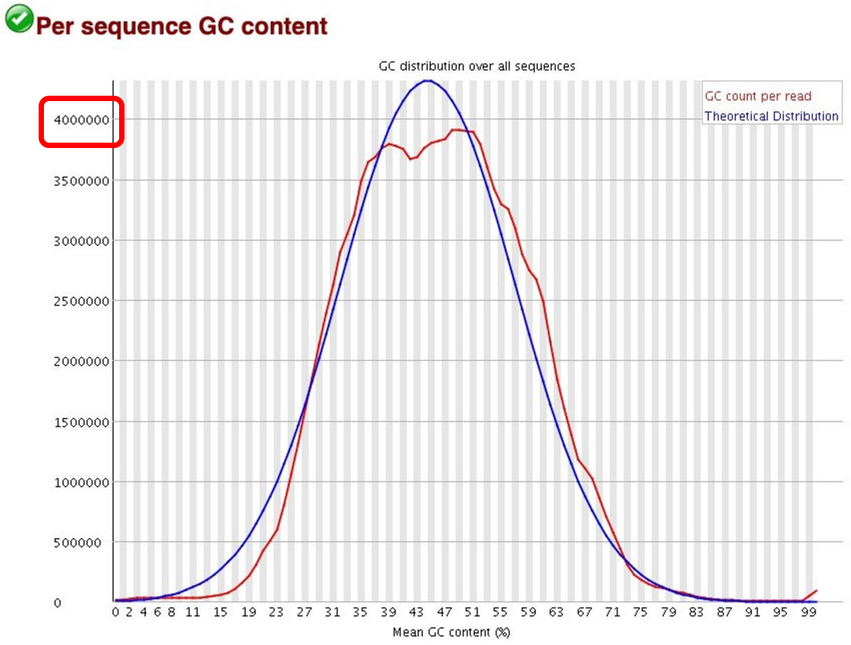


Fail
Organoid sample snRNA-seq

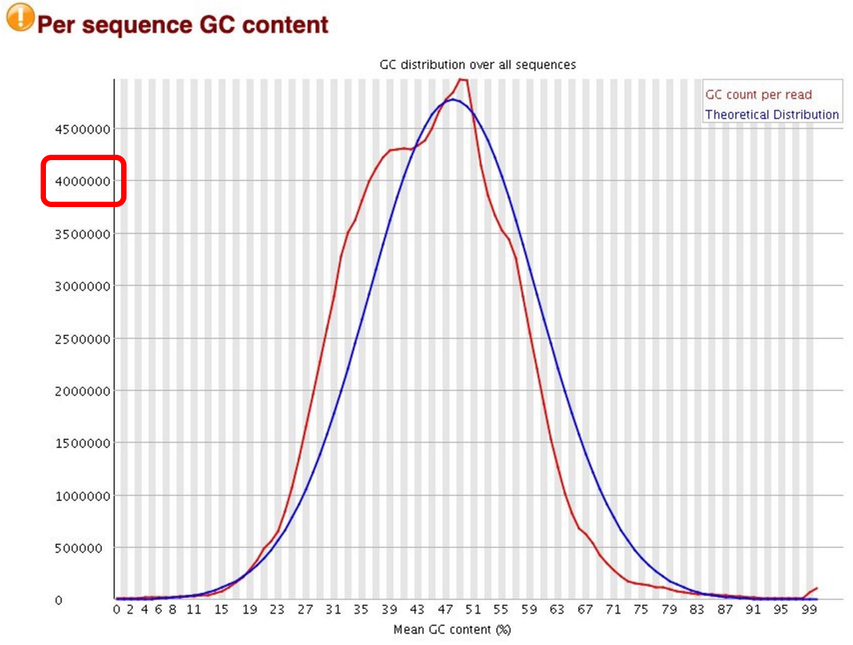


Per sequence GC content indicates if distribution of GC content is non-normal

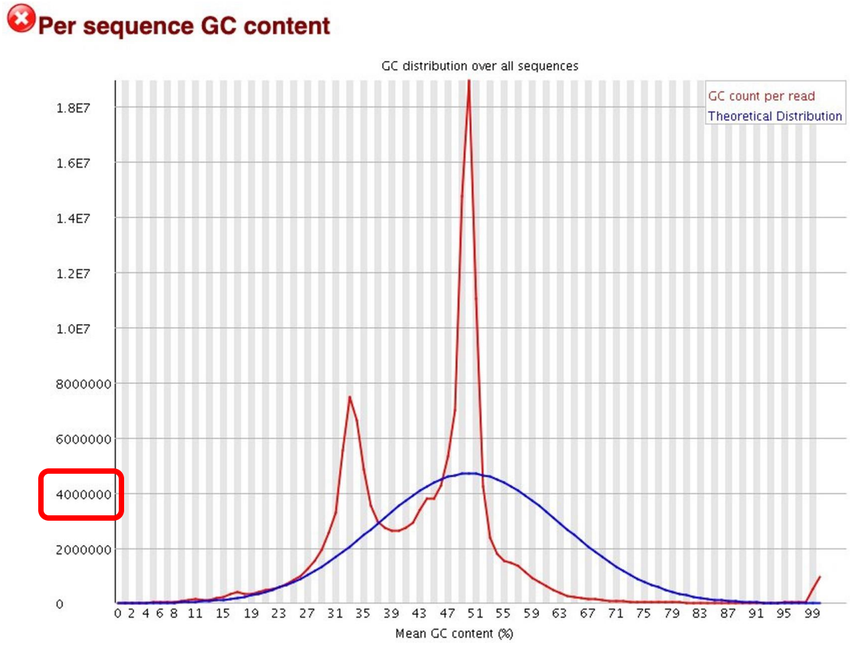
Gold standard
10x Genomics scRNA-seq



Salvageable?
Patient sample snRNA-seq

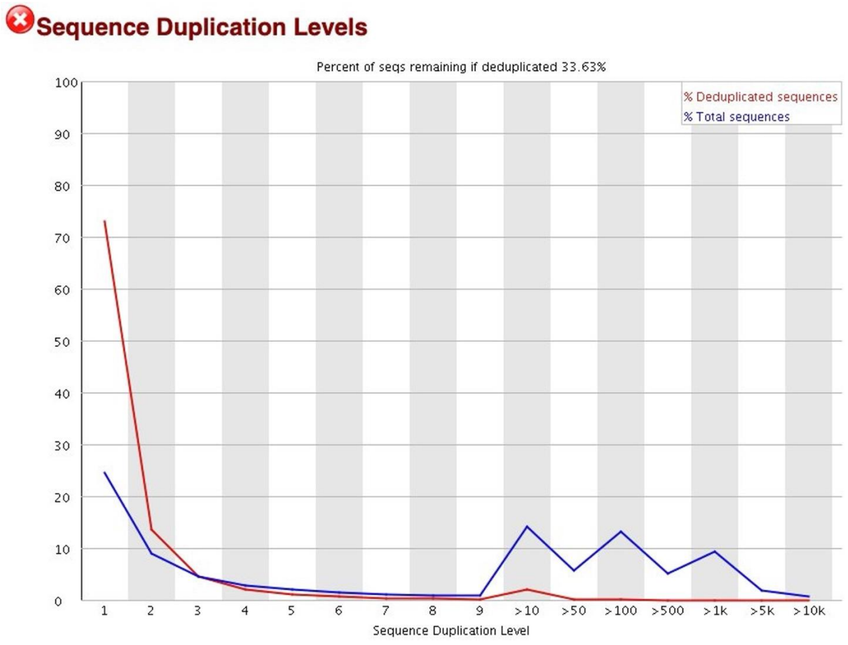


Fail
Organoid sample snRNA-seq

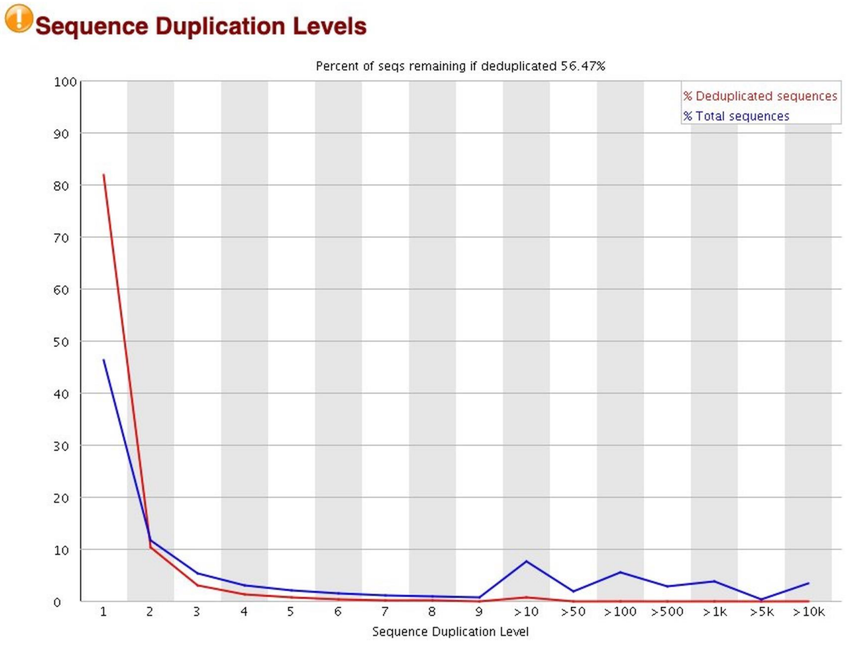


Sequence duplication levels indicate possible enrichment bias

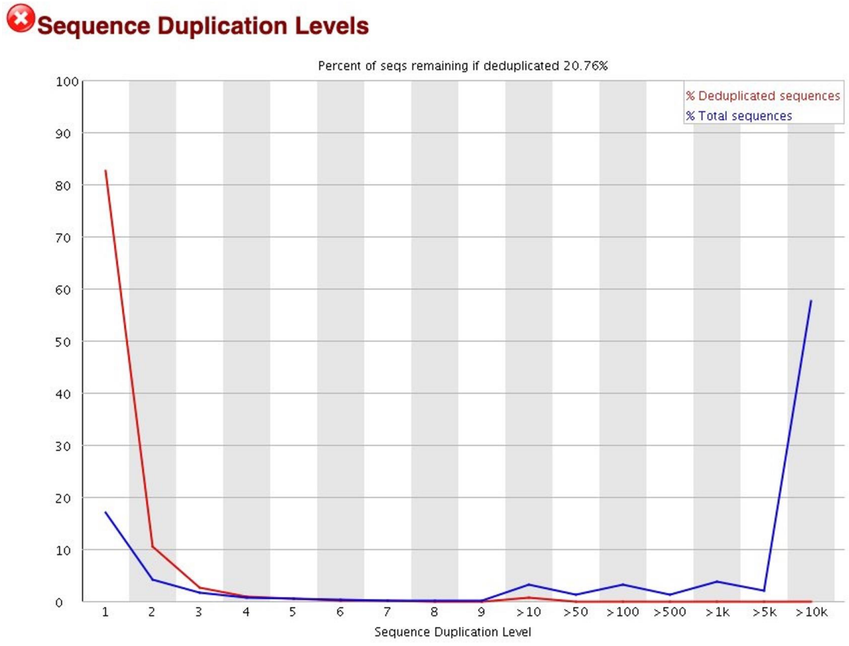
Gold standard
10x Genomics scRNA-seq



Salvageable?
Patient sample snRNA-seq



Fail
Organoid sample snRNA-seq



Overrepresented sequences provides a list of sequences that appear more often than expected

Gold standard
10x Genomics scRNA-seq

Overrepresented sequences
No overrepresented sequences

Salvageable?
Patient sample snRNA-seq

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAG	795438	0.6026238796373505	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGAAAAAAAAAAAAAAAAAAAA	661294	0.5009963766640544	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGTCAGATGTGTATAAGAGAC	439409	0.33289628648313074	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGCAGTGGTATCAACGCAGAG	245919	0.18630825011696397	No Hit
GCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAG	169701	0.12856548844578458	No Hit
GTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTAC	147090	0.11143539340068978	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	145015	0.10986337326807417	No Hit

7 overrepresented sequences, ~2% of total sequences

Fail
Organoid sample snRNA-seq

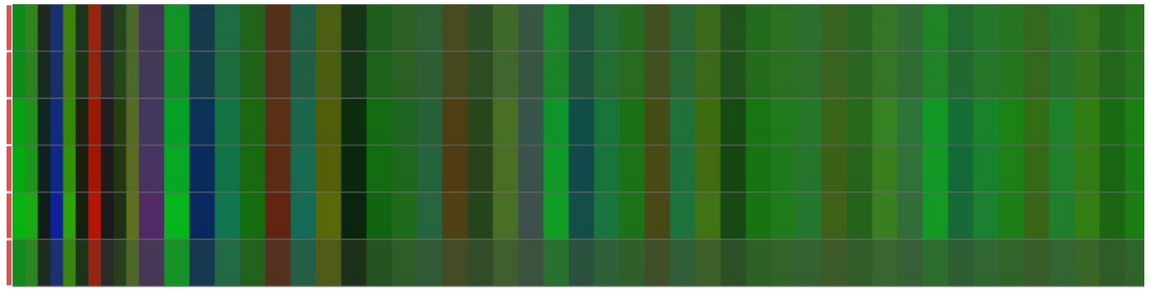
Overrepresented sequences

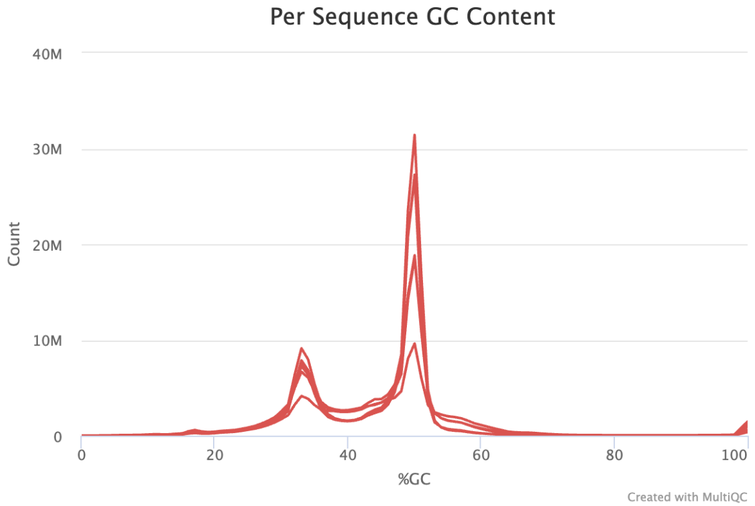
Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAG	25939446	18.254145010928653	No Hit
GTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTAC	5171528	3.6393152745080926	No Hit
GCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAG	3863598	2.718896855234839	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGAAAAAAAAAAAAAAAAAAAA	2649906	1.8647957396364558	No Hit
CAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGT	2588501	1.8215837229111922	No Hit
AGCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGA	2165226	1.5237160186625809	No Hit
GGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTACAT	1924853	1.3545603783026459	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGTGGTATCAACGCAGAGTAC	1528207	1.0754320730179143	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGTCAGATGTGTATAAGAGAC	1335247	0.9396419786069237	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1330865	0.9365582711353805	No Hit
ATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTACATGGG	1305581	0.9187653775455822	No Hit
AGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTA	1254592	0.882883323628076	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGAAAAAAAAAAAAAAAAAAA	1034366	0.7279055596782688	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGCAGTGGTATCAACGCAGAG	884684	0.6225711229472057	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGAGCAGTGGTATCAACGCAGA	878123	0.6179540063975036	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGAAAGCAGTGGTATCAACGCA	817202	0.5750825908626159	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGTGGTATCAACGCAGAGTACA	747318	0.5259037198125658	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGAAGCAGTGGTATCAACGCA	733231	0.5159903954968131	No Hit
GTGGTATCAACGCAGAGTACATGGGAAAAAAAAAAAAAAAAAAAAAAAA	667723	0.4698910095894998	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGTATCAACGCAGAGTACATG	636939	0.4482276479272489	No Hit
GTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTACATG	534040	0.37581541262046764	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGCAGTGGTATCAACGCAGAGT	485514	0.34166662467795245	No Hit
GCAGTGGTATCAACGCAGAGTACATGGGAAAAAAAAAAAAAAAAAAAAA	478518	0.33674338928979275	No Hit
GTGGTATCAACGCAGAGTACATGGGGTGGTATCAACGCAGAGTACATGGG	454924	0.32013978498043894	No Hit
TATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTACATGG	407916	0.28705924622592066	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGAAAAGCAGTGGTATCAACGC	377545	0.26568652152983757	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGAAGCAGTGGTATCAACGCAGA	356291	0.25072962545494537	No Hit

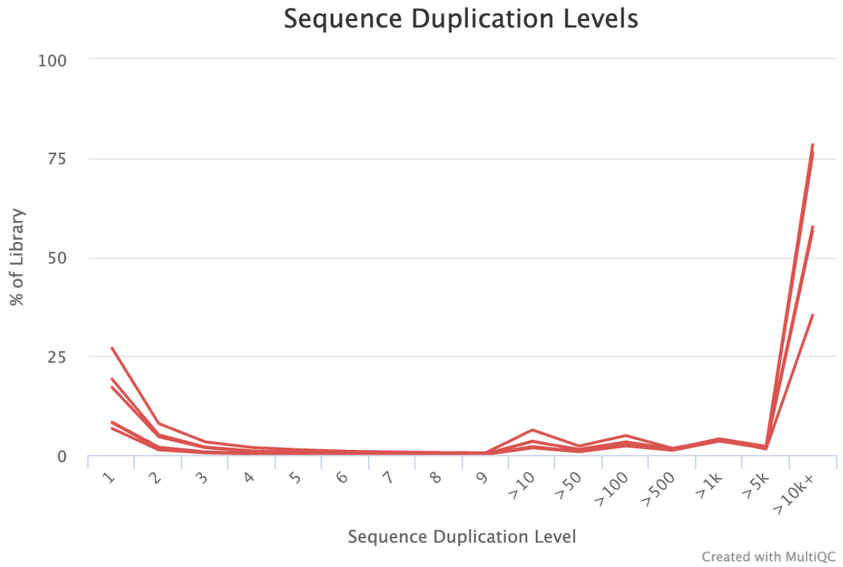
60 overrepresented sequences, ~47% of total sequences

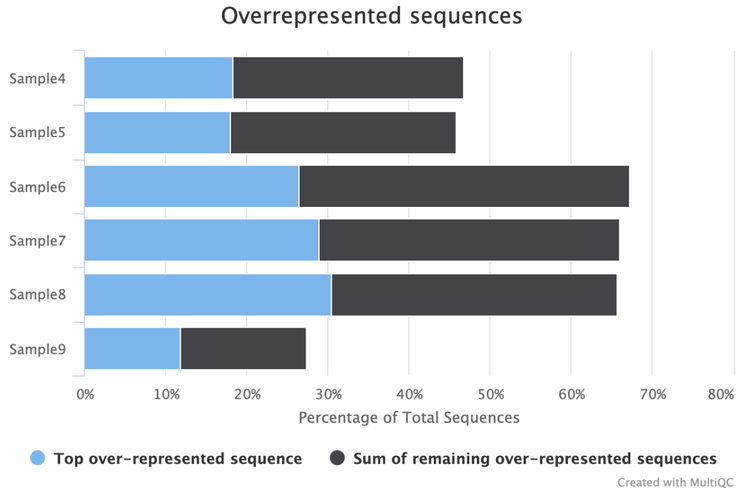
These are examples of good sequencing of bad or heavily degraded samples

Per Base Sequence Content

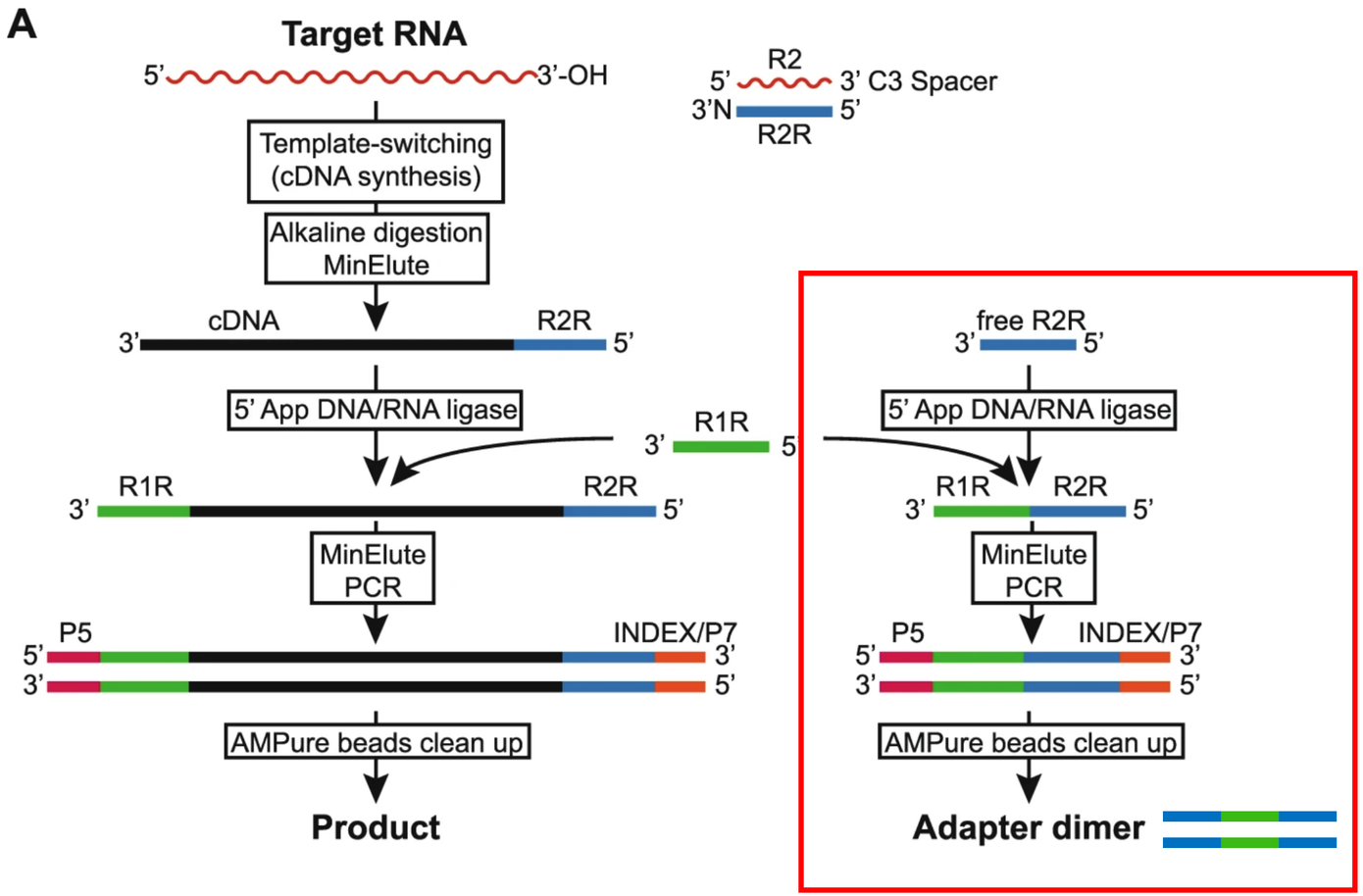








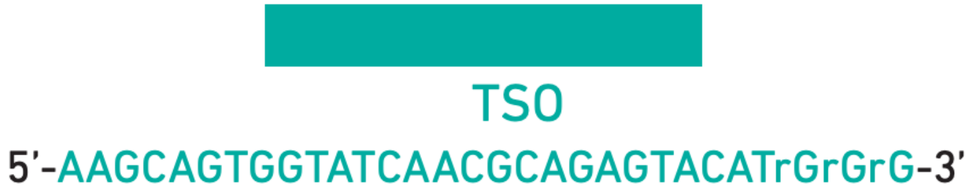
Adapter dimers can pass early QC steps when dimerization produces long sequences



30bp adapters (90 bases total)

Xu, H., Yao, J., Wu, D.C. et al. DOI: 10.1038/s41598-019-44457-z

Overrepresented sequences show repetition of template switching oligo (TSO) sequence



Sequence
AAGCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAG
AGCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGA
GCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAG
GTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGTAC
CAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAGAGT
AAGCAGTGGTATCAACGCAGAGTACATGGGAAAAAAAAAAAAAAAAAAAA
AAGCAGTGGTATCAACGCAGAGTACATGGG*GTGGTATCAACGCAGAGTAC*

Quality control of sequencing alignment with CellRanger

6,257

Estimated Number of Cells

51,153

Mean Reads per Cell

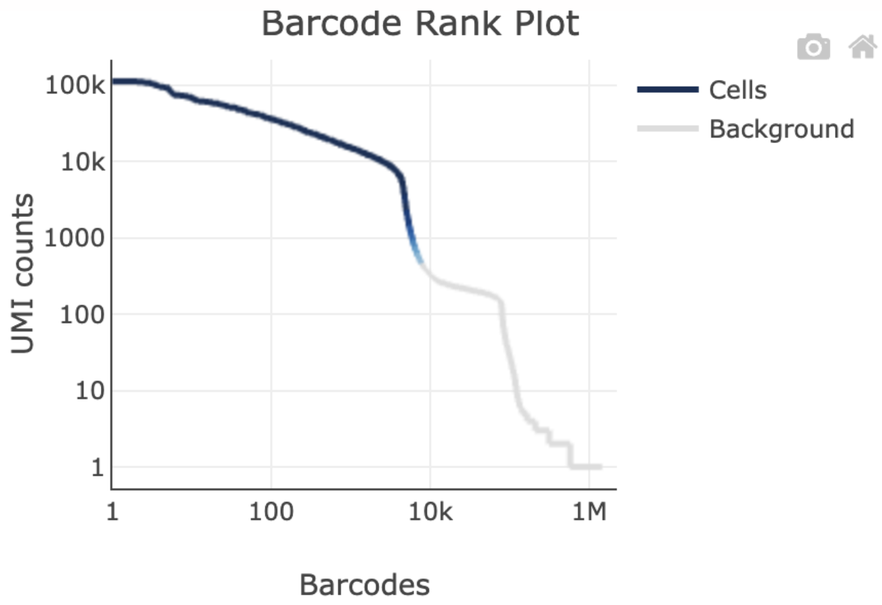
3,707

Median Genes per Cell

Sequencing

Number of Reads	320,065,331
Number of Short Reads Skipped	0
Valid Barcodes > 75%	94.8%
Valid UMIs > 75%	99.9%
Sequencing Saturation depends on sequence depth and sample complexity	46.3%
Q30 Bases in Barcode > 85%	96.2%
Q30 Bases in RNA Read > 85%	93.3%
Q30 Bases in UMI > 85%	95.8%

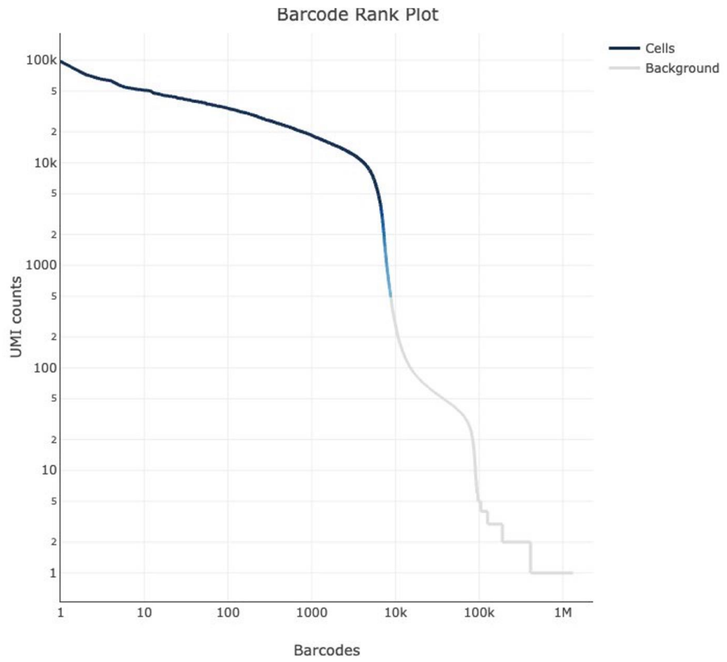
Cells

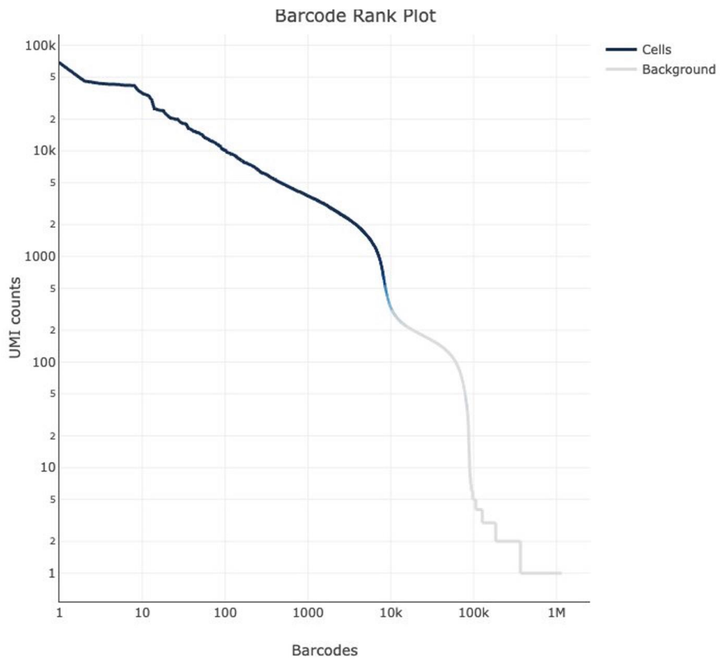


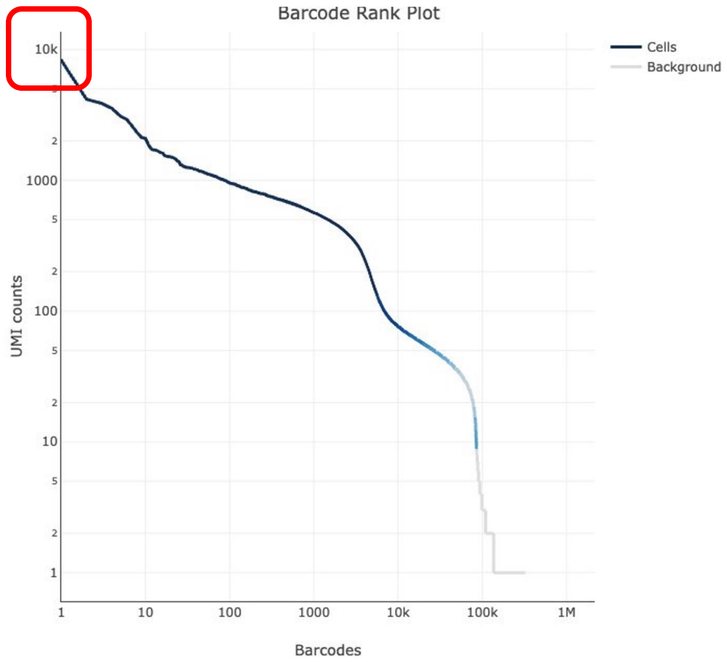
Estimated Number of Cells 500 – 10,000	6,257
Fraction Reads in Cells > 70%	78.4%
Mean Reads per Cell > 20,000	51,153
Median Genes per Cell > 1,000	3,707
Total Genes Detected depends on sequence depth and sample complexity	31,601
Median UMI Counts per Cell depends on sequence depth and sample complexity	8,828

Mapping

Reads Mapped to Genome > 85%	87.1%
Reads Mapped Confidently to Genome > 80%	80.7%
Reads Mapped Confidently to Intergenic Regions < 10%	9.5%
Reads Mapped Confidently to Intronic Regions < 30%	47.8%
Reads Mapped Confidently to Exonic Regions > 30%, ~60%	23.4%
Reads Mapped Confidently to Transcriptome > 30%, ~60%	48.3%
Reads Mapped Antisense to Gene < 10%	22.0%





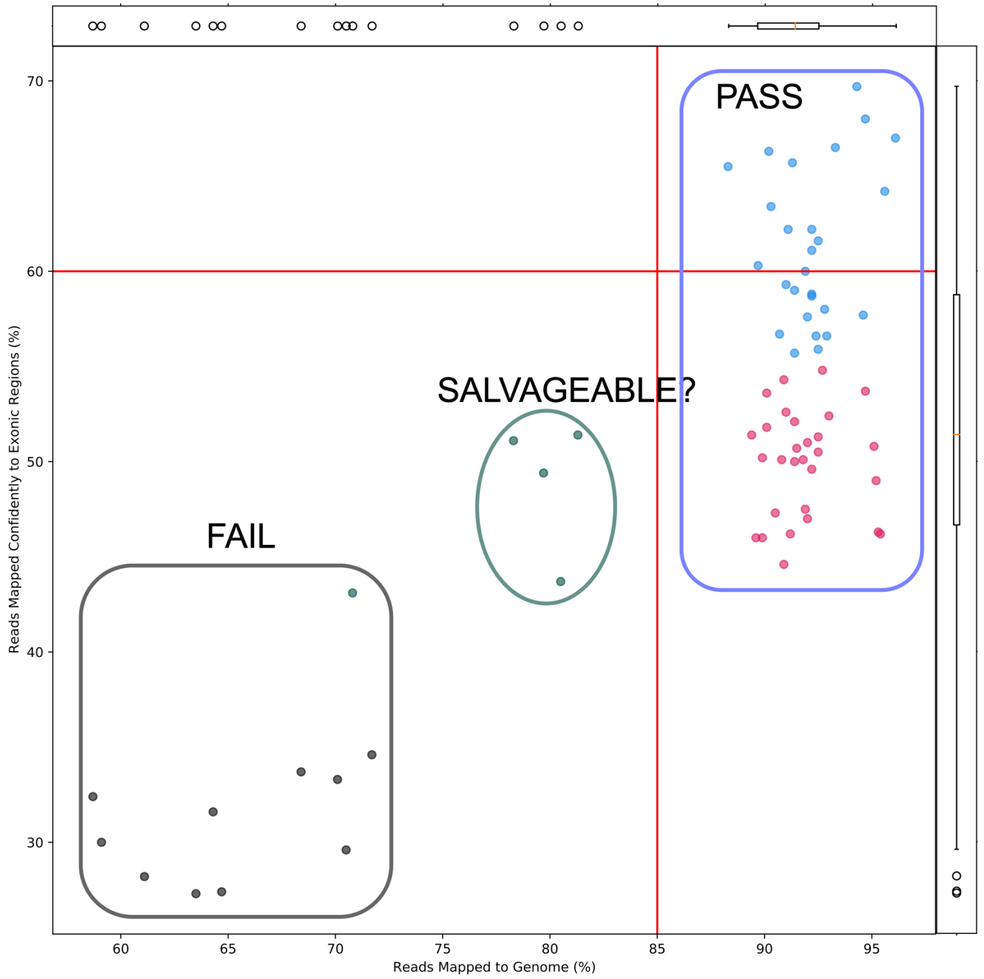


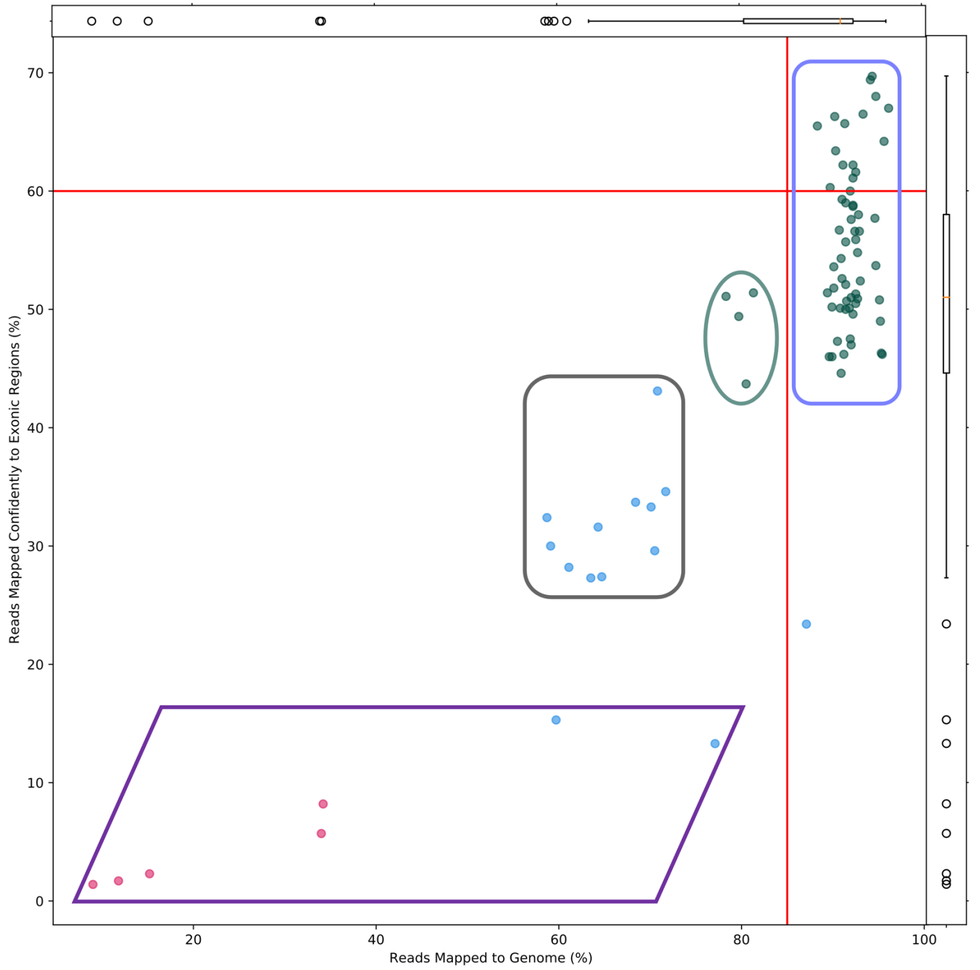
	Good Sample
Number of Reads	293,132,076
Q30 Bases in RNA Read	94.90%
Mapped to Genome	92.70%
Mapped Confidently to Genome	89.30%
Mapped Confidently to Exonic Regions	50.90%
Mapped Confidently to Transcriptome	47.50%
Number of Cells after Filtering	6,510

Sample with Concerns
209,364,338
90.30%
59.70%
53.30%
15.30%
19.80%
6,841

Outlier Sample
316,003,364
87.60%
9.00%
6.40%
1.40%
2.20%
2,217

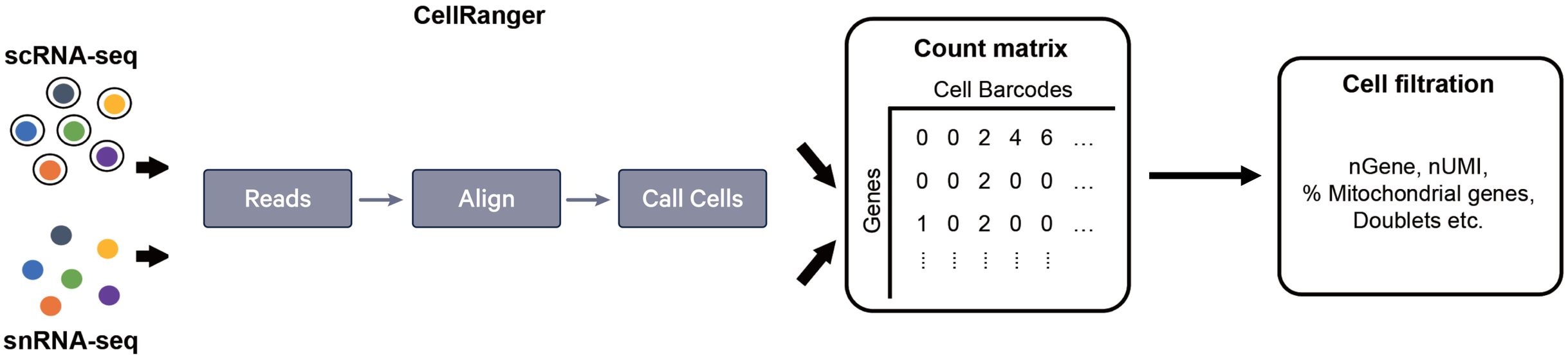
Example of read mapping metrics for range of samples





Filtering low quality cells with Seurat

Count matrix (aka gene-barcode matrix) from CellRanger is input into Seurat for cell-level filtering

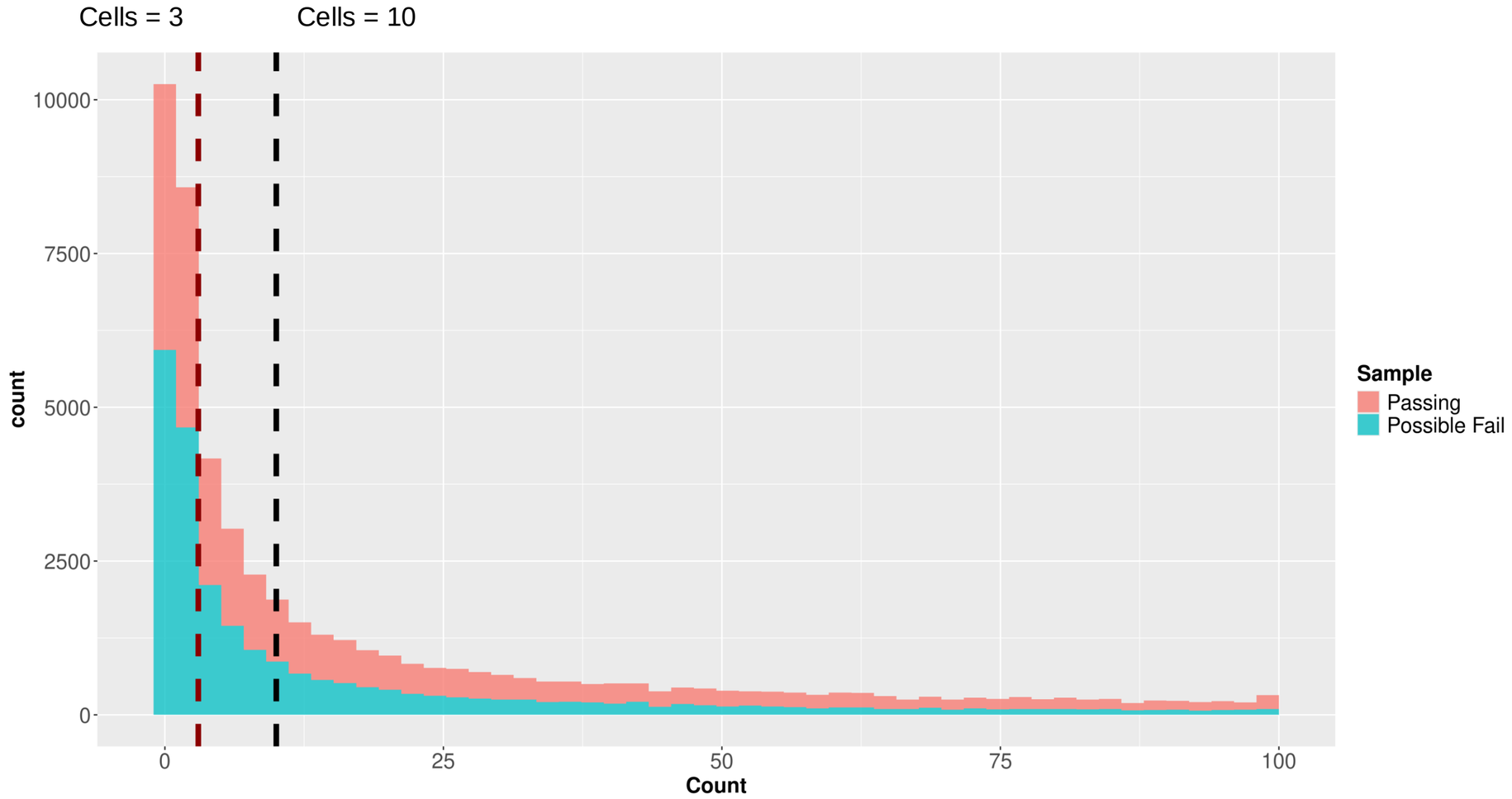


Kim N, Kang H, Jo A, Yoo SA, Lee HO. DOI: 10.4132/jptm.2022.12.19

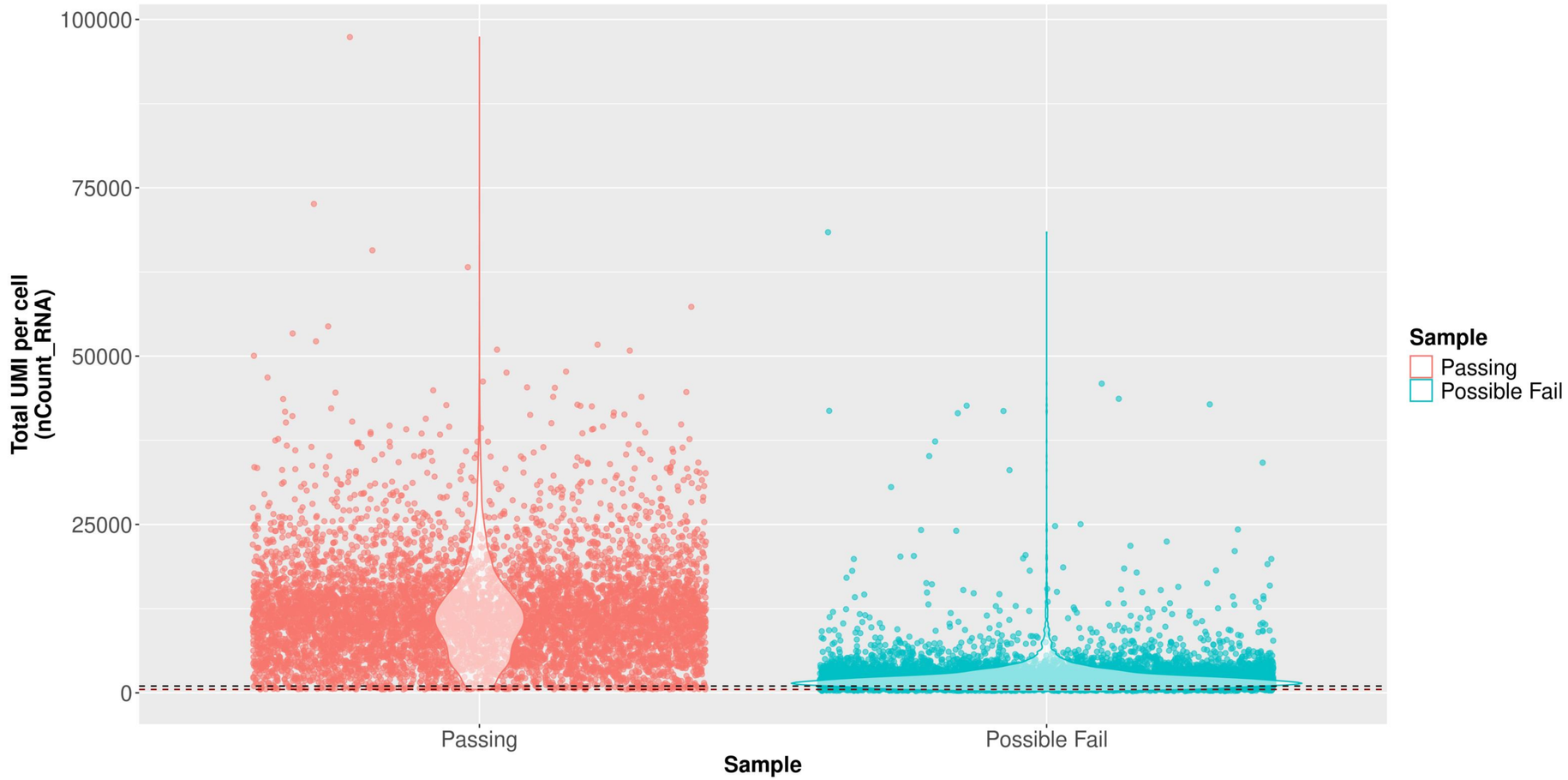
Baseline values for cell-level filtering we use (may change based on each specific experiment)

Technology	Cut-offs	Filter 1	Filter 2 (Outliers, Optional)
		Cells per gene	
scRNA	Min	10	
	Max	-	
snRNA	Min	10	
	Max	-	

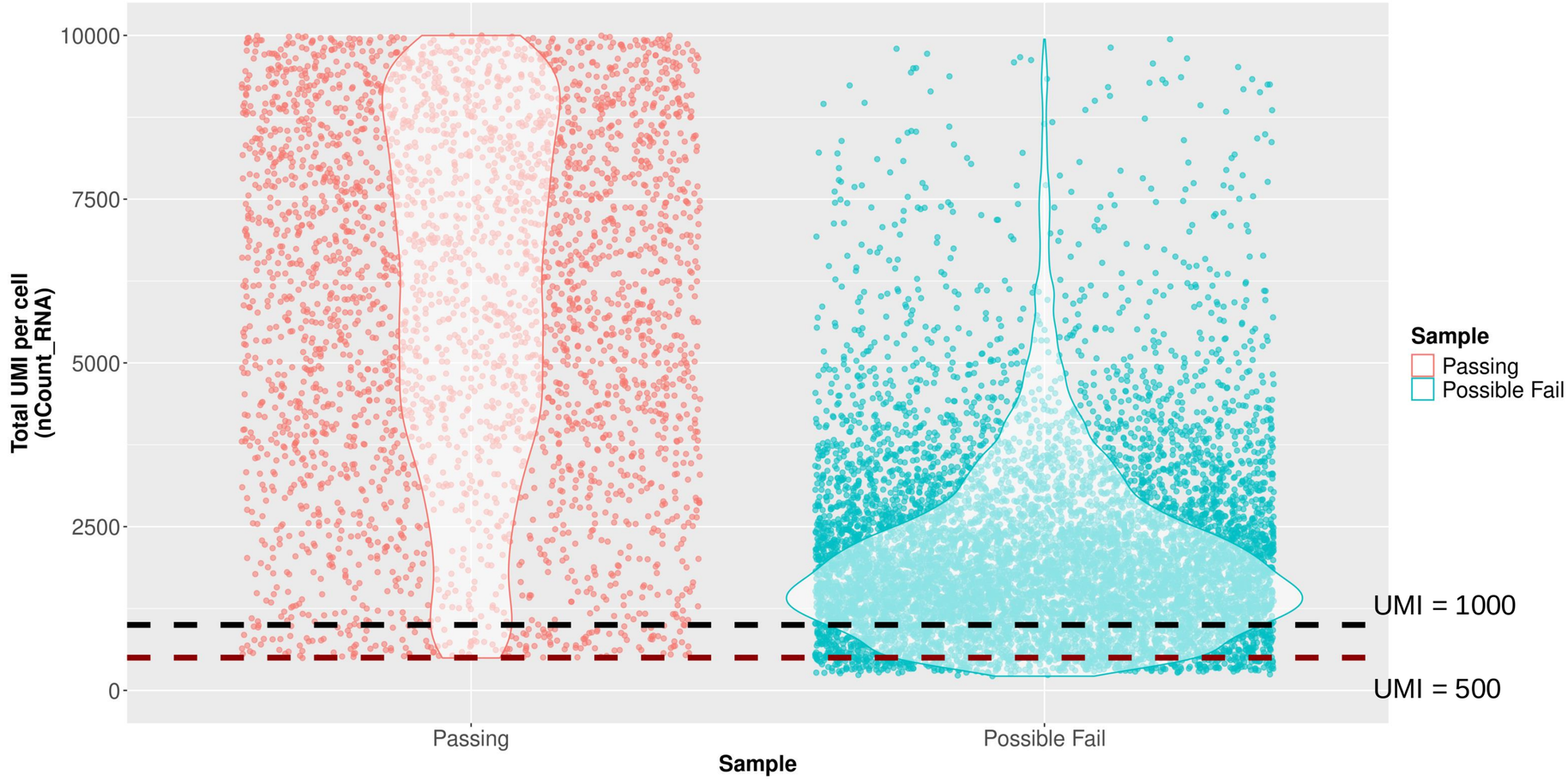
Cells per gene threshold can be used to remove lowly detected genes



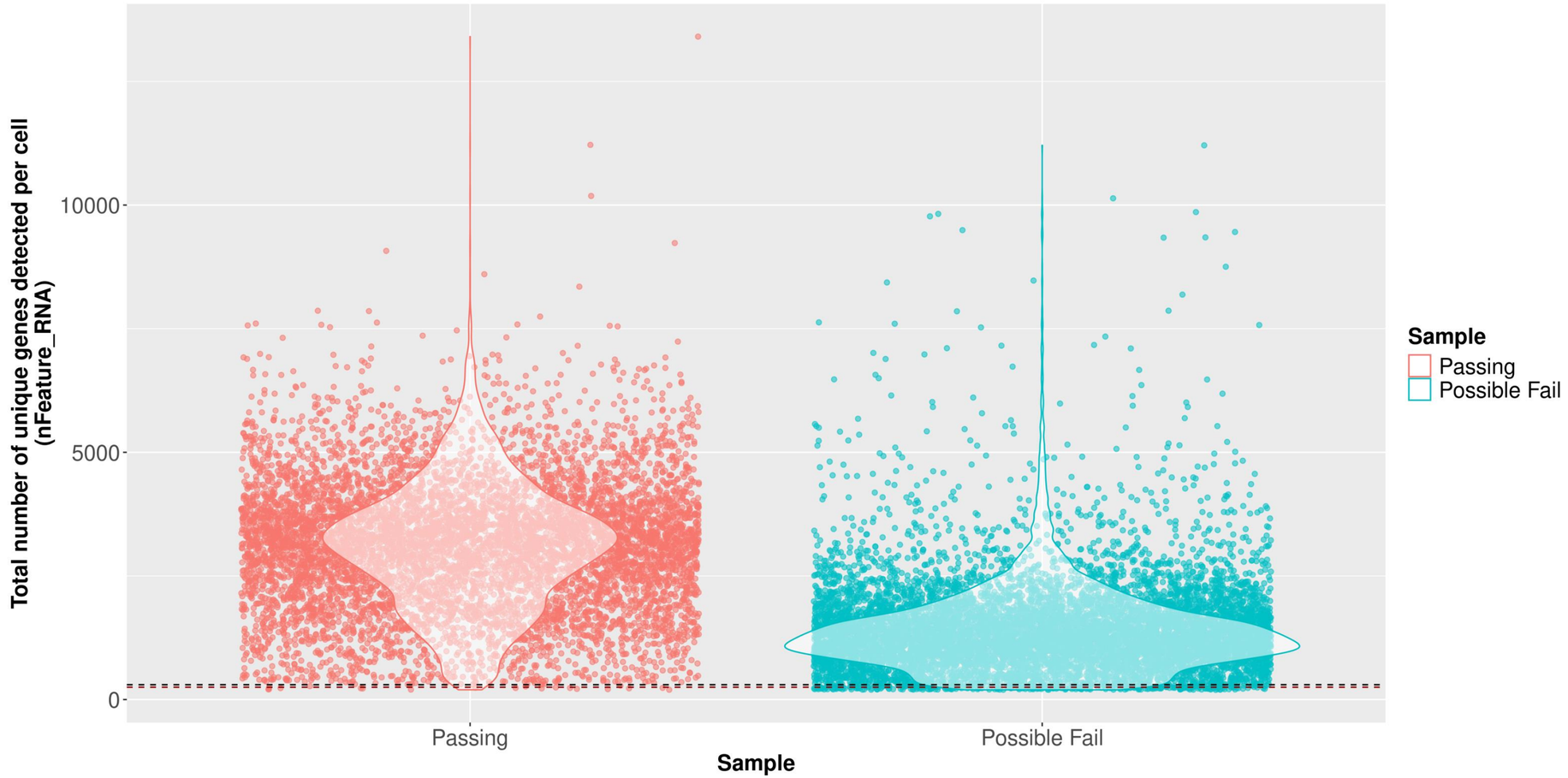
UMI counts (transcripts) per cell can be used to filter out “empty” cells or poorly sequenced cells



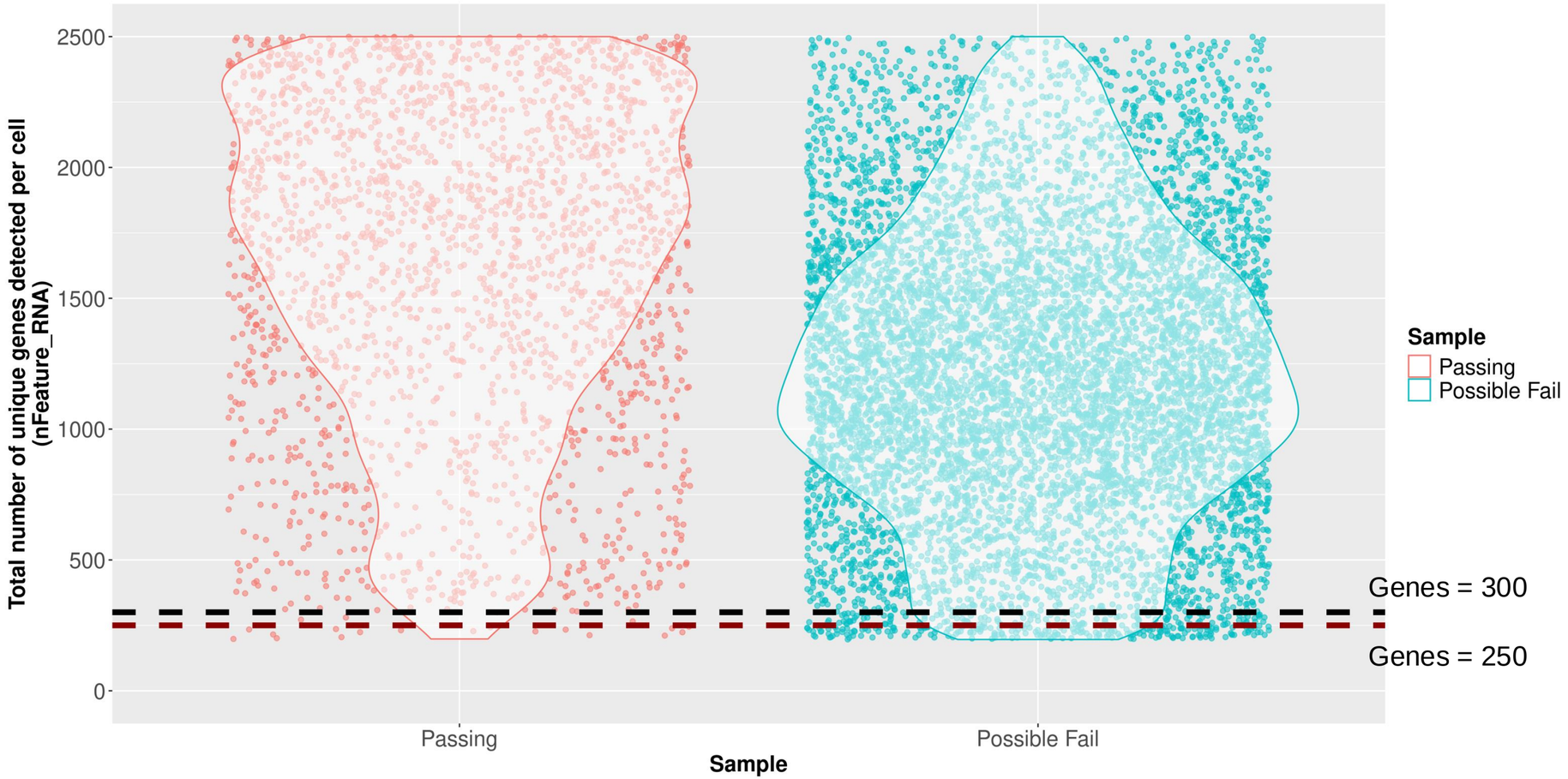
UMI counts (transcripts) per cell can be used to filter out “empty” cells or poorly sequenced cells



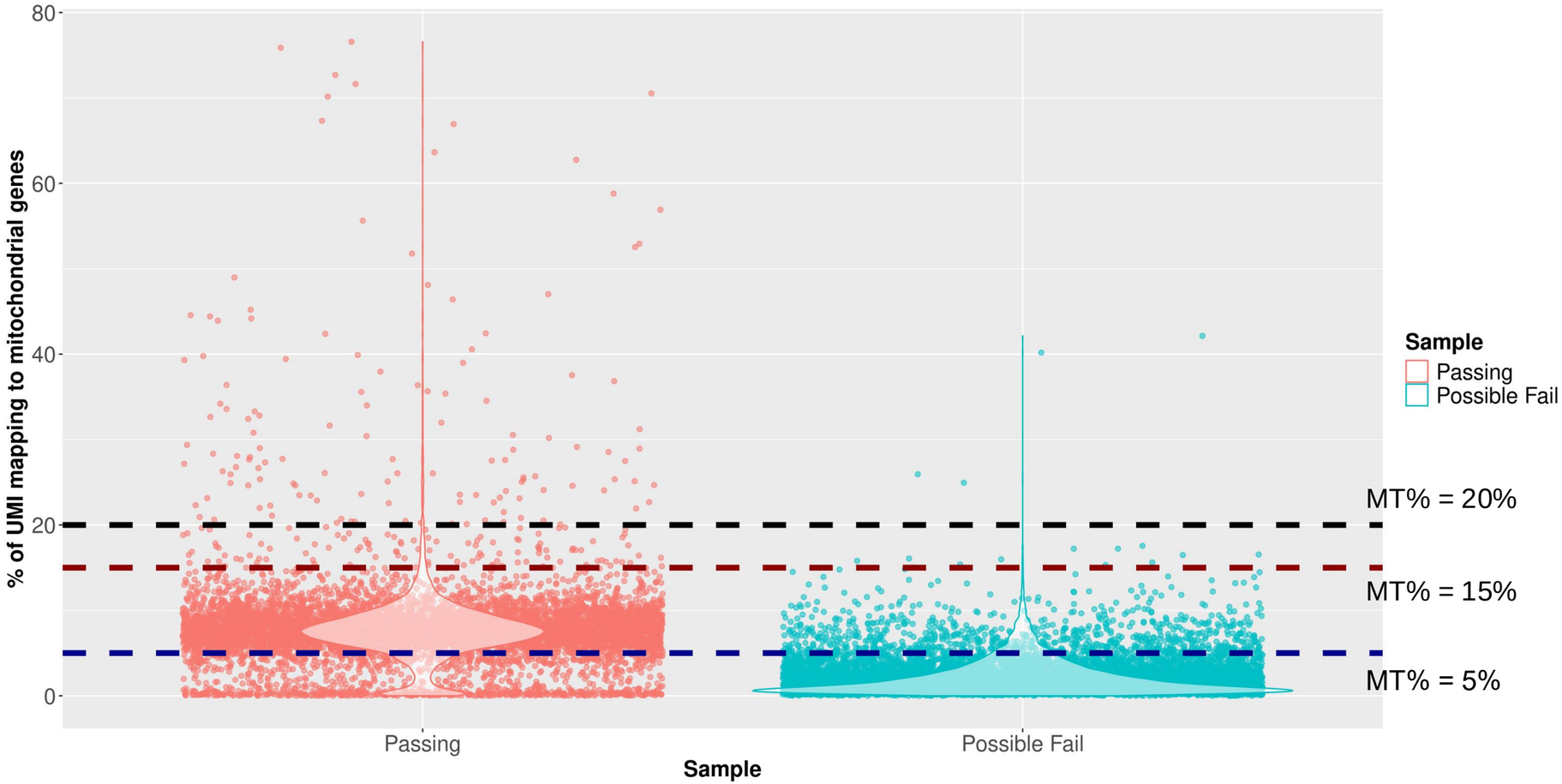
Genes (features) detected per cell can be dependent on cell type



Genes (features) detected per cell can be dependent on cell type



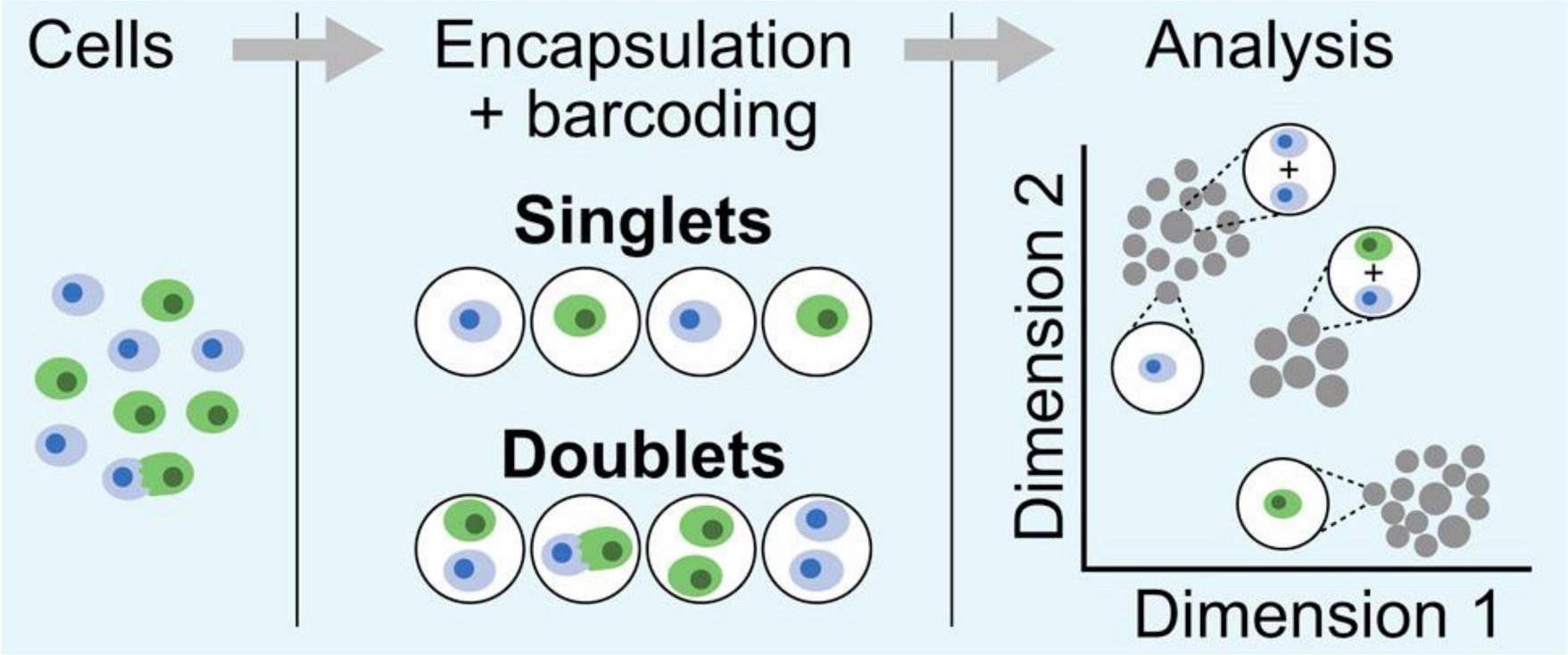
Mitochondrial percentage parameter can be dependent on the sample type, cell type



Cell-level filtering values are ultimately based on each specific experiment

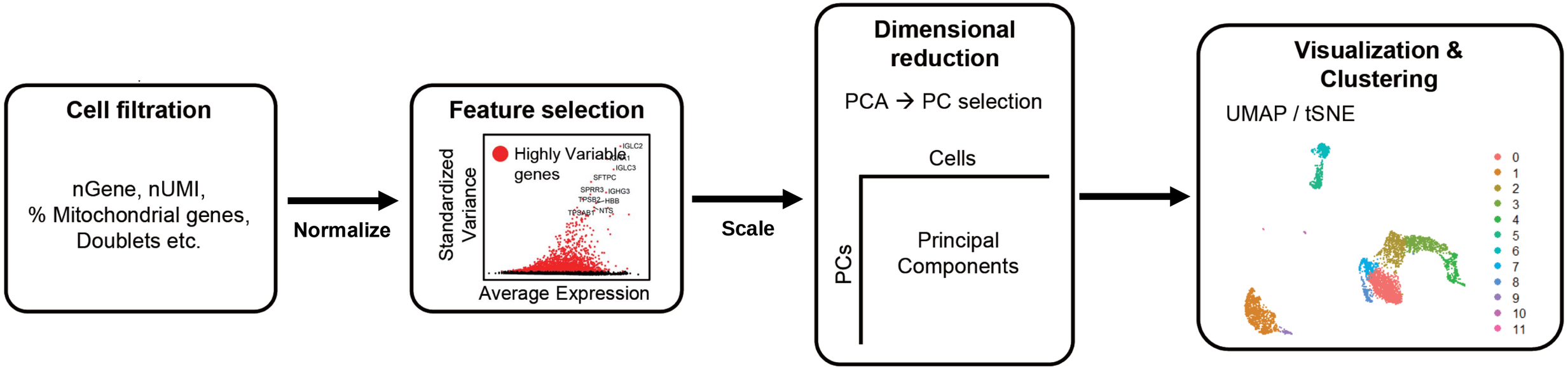
```
# Filter out low quality cells using selected thresholds - these will change with experiment
filtered_seurat <- subset(x = merged_seurat,
                          subset= (nUMI >= 500) &
                            (nGene >= 250) &
                            (mitoRatio < 0.20))
```

Cell-level filtering can also include doublet detection



Wolock SL, Lopez R, Klein AM.. DOI: 10.1016/j.cels.2018.11.005

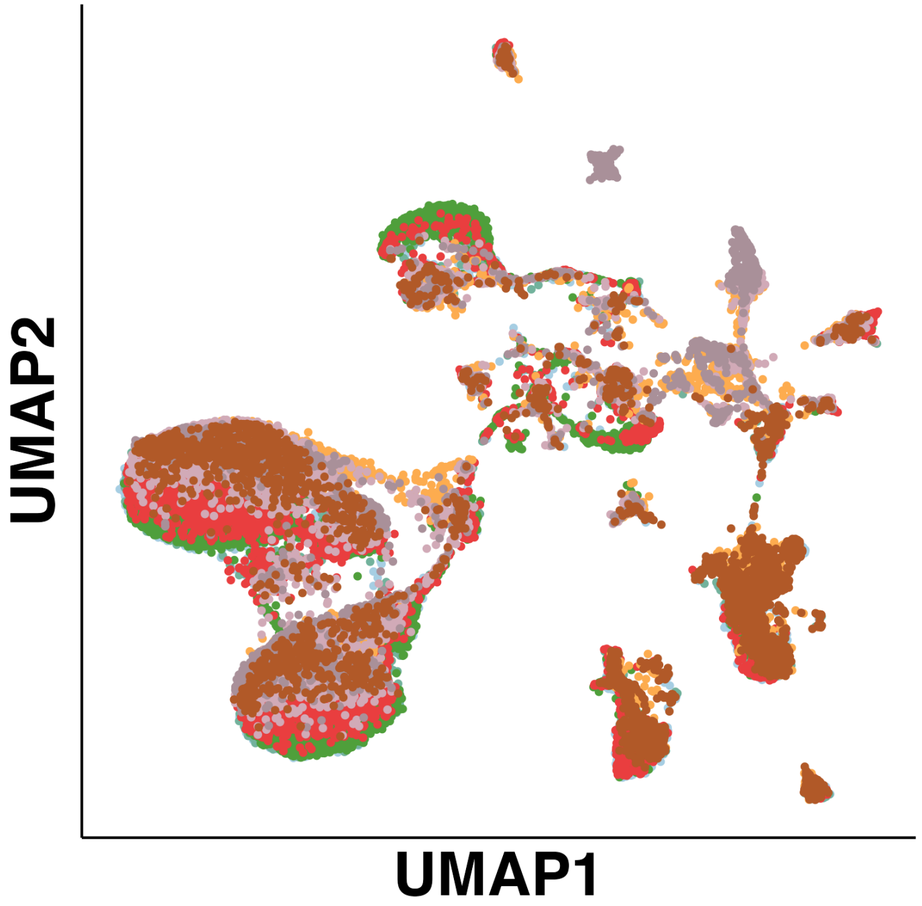
After cell-level filtering, downstream analysis typical begins with dimensionality reduction and clustering

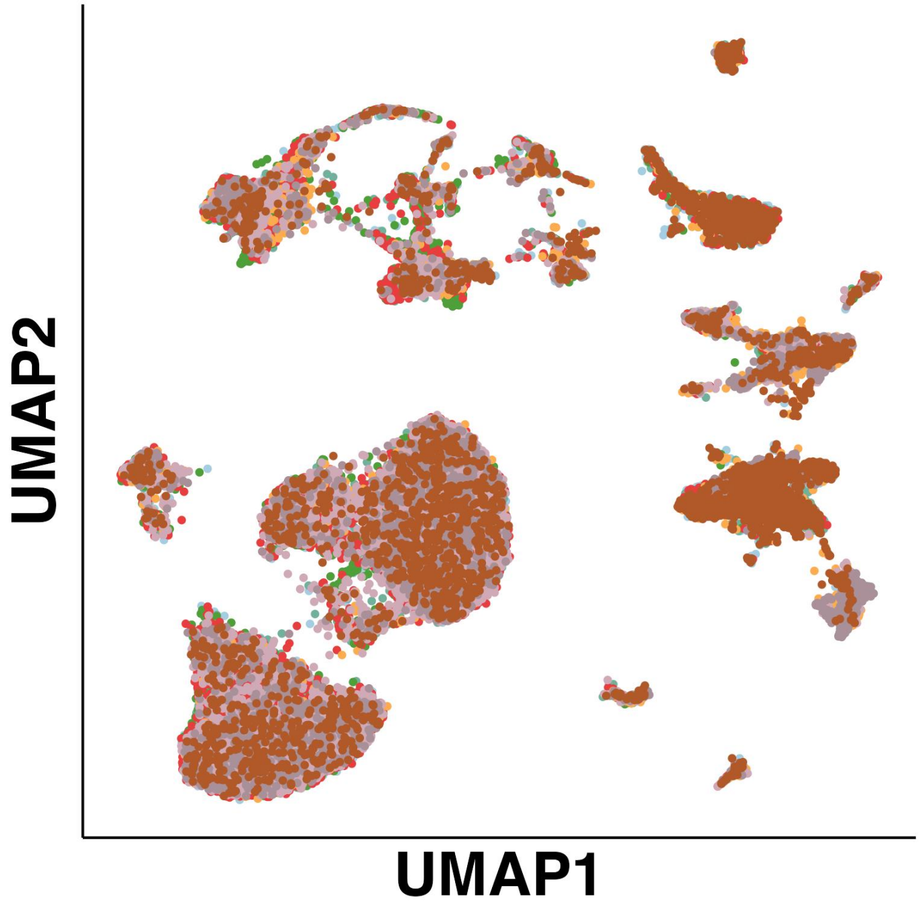


Kim N, Kang H, Jo A, Yoo SA, Lee HO. DOI: 10.4132/jptm.2022.12.19

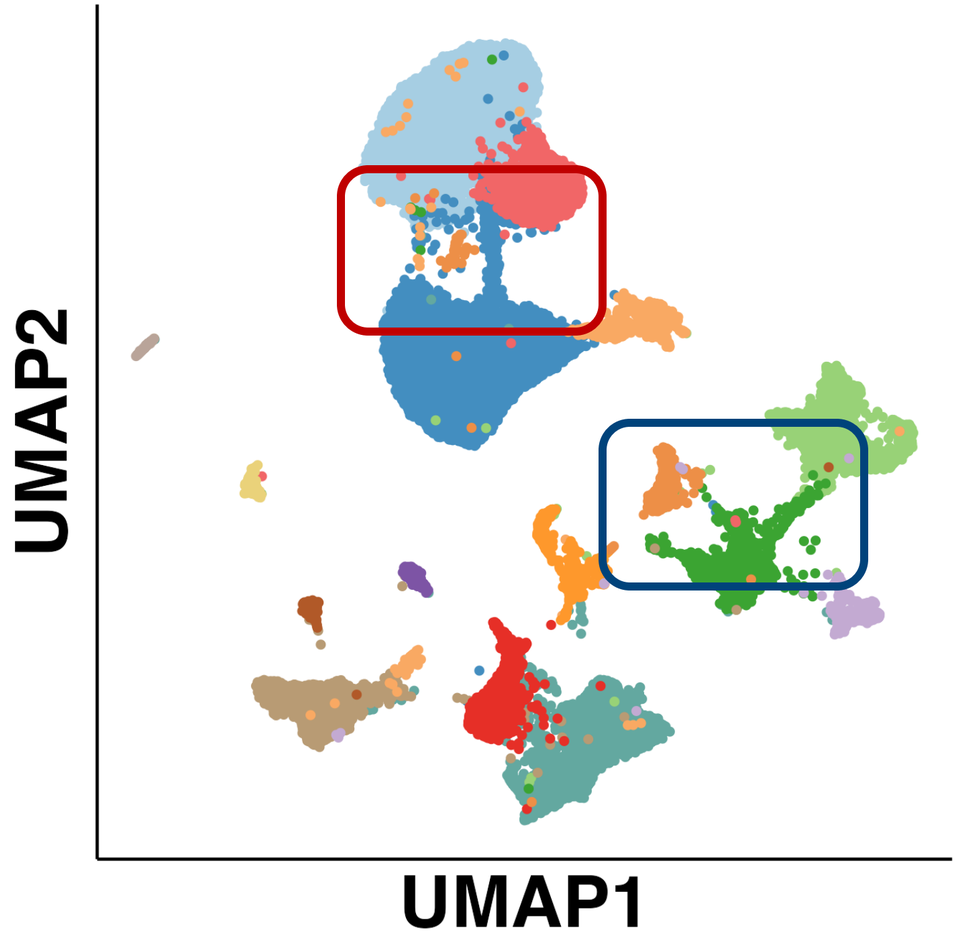
Additional processing steps before “finalizing” dataset for analysis (Harmony, SoupX)

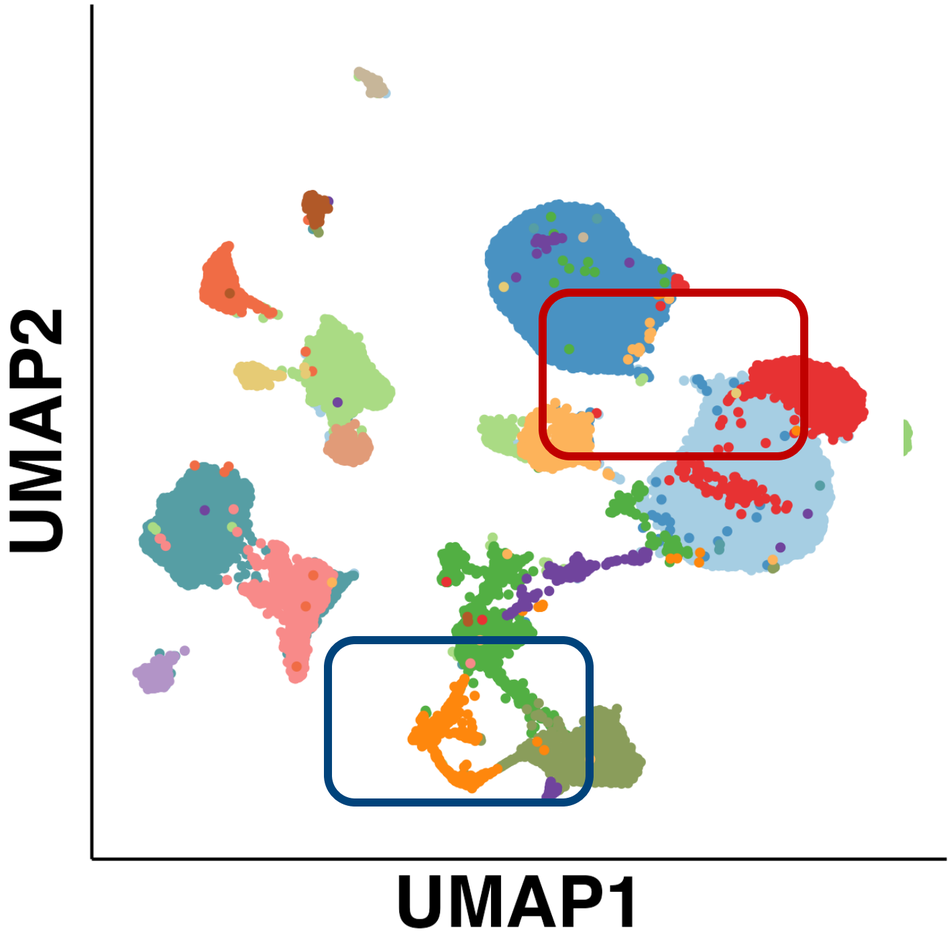
When analyzing multiple samples, more sophisticated integration methods may be used (e.g. Harmony)





Some samples may need to be corrected for ambient RNA contamination (e.g. SoupX)





General analysis steps with Seurat (and other tools)

Once preparation of ”final” gene-barcode matrix, clustering is completed, common analysis steps include:

- Cell typing
  - Based on curated marker expression, DE genes
  - Module scoring method
  - Automated methods comparing to reference expression profiles (e.g. SingleR)
  - Reference label transfer (mapping your data onto a labeled reference data set containing cell types you expect in your data)
- Differential expression
  - Comparing clusters or cell types
  - Comparing conditions or phenotypes
  - Comparing timepoints
- Sub-clustering of specific clusters, subsets of cells
  - E.g. if you want to separate out more subtly cell types or subtypes
- Pseudotime or trajectory analysis
  - Position cells across spectrum of time or differentiation stage (for example)
  - See how expression profiles changes across time/stage
  - See how distribution of cell types changes over time/stage

Questions?

Break for lunch

Visualization of sc/snRNA-seq with R Shiny (ShinyCell)

Exploring a retinal scRNA-seq dataset using R Shiny app

- Looking at core regulatory circuit super enhancer (CRC-SE) for Vsx2
- Vsx2 expressed in retinal progenitor cells
- Vsx2 expressed in differentiated bipolar cells, Muller glia
- Deleted 32kb region of CRC-SE
- Deleted subregions to identify specific effects
- Performed multi-omic analysis
- Including scRNA-seq of retinal cell populations
- https://vsx2-deletion.stjude.org/

Identification of a modular super-enhancer in murine retinal development

Victoria Honnell[1,2], Jackie L. Norrie[1], Anand G. Patel[1], Cody Ramirez[1], Jiakun Zhang[1], Yu-Hsuan Lai[1], Shibiao Wan[3] & Michael A. Dyer[1✉]

Super-enhancers are expansive regions of genomic DNA comprised of multiple putative enhancers that contribute to the dynamic gene expression patterns during development. This is particularly important in neurogenesis because many essential transcription factors have complex developmental stage- and cell-type specific expression patterns across the central nervous system. In the developing retina, Vsx2 is expressed in retinal progenitor cells and is maintained in differentiated bipolar neurons and Müller glia. A single super-enhancer controls this complex and dynamic pattern of expression. Here we show that deletion of one region